УДК 004.934

***В.В. Робейко, М.М. Сажок***
Міжнародний науково-навчальний центр інформаційних технологій та систем «КіберМова», м. Київ, Україна
Україна, 03680, просп. Акад. Глушкова, 40, МСП, м. Київ, {valya.robeiko, sazhok}@gmail.com

# Розпізнавання спонтанного мовлення на основі акустичних композитних моделей слів у реальному часі

***V.V. Robeiko, M.M. Sazhok***
*Speech Science and Technology Department, International Research and Training Center of Information Technologies and Systems «CyberMova», Kyiv, Ukraine*
*Ukraine, 03680, Acad. Glushkov Ave., 40, MSP, Kyiv,*{valya.robeiko, sazhok}@gmail.com

# *Real-Time Spontaneous Speech Recognition Based on Word Acoustic Composite Models*

***В.В. Робейко, М.М. Сажок***
Международный научно-учебный центр информационных технологий и систем «КиберМова»; г. Киев, Украина
Украина, 03680, пр. Акад. Глушкова, 40, МСП, г. Киев, {valya.robeiko, sazhok}@gmail.com

# Распознавание спонтанной речи на основе акустических композитных моделей слов в реальном времени

У статті розглядається реалізація методів і алгоритмів розпізнавання злитого мовлення на основі композиції слів із акустичних генеративних моделей фонем. Аналізуються аспекти оцінки параметрів математичних моделей акустичної та лінгвістичної складових системи розпізнавання та перетворення графем на фонеми, що поєднує обидві ці складові. Окрема увага приділяється прогнозуванню наголосів у словах та врахуванню ознак спонтанності. Базова експериментальна система розпізнавання злитого (у тому числі спонтанного) мовлення в реальному часі оперує словником до ста тисяч слів та дає змогу набирати текст під диктування. Аналізуються перспективи подальшого розширення словника та вдосконалення процедур оцінки параметрів моделей, обговорюються ергономічні питання.
**Ключові слова:** розпізнавання мовлення, спонтанне злите мовлення, генеративна модель, реальний час.

This paper describes implementation of methods and algorithms for the automatic speech recognition based on word composition proceeding from acoustic phoneme models. Such a design of the speech-to-text decoder is conventional and most productive for Western languages. The aim is to explore this approach applied to the Ukrainian language that is highly inflective with relatively free word order. We use data-driven methods to estimate parameters for both acoustic and linguistic components of the mathematical model. The grapheme-to-phoneme conversion procedure takes into account word stress issue and spontaneous continuous speech features. The basic speech-to-text system is able to operate a 100k vocabulary in real-time. The prospective of dictionary and domain extension, parameter estimation improvement and ergonomic issues are discussed.
**Key words:** Speech recognition, spontaneous continuous speech, generative model, real-time.

Рассматривается реализация методов и алгоритмов распознавания слитной речи на основе композиции слов из акустических генеративных моделей фонем. Анализируются аспекты оценки параметров математических моделей акустической и лингвистической составляющей системы распознавания и

преобразования графем в фонемы, объединяющей обе эти составляющие. Отдельное внимание уделяется прогнозированию ударений в словах и учету признаков спонтанности. Базовая экспериментальная система распознавания слитной (в том числе спонтанной) речи в реальном времени оперирует словарем до ста тысяч слов, и позволяет набирать текст под диктовку. Анализируются перспективы дальнейшего расширения словаря и совершенствования процедур оценки параметров моделей, обсуждаются эргономические вопросы.

**Ключевые слова:** распознавание речи, спонтанная слитная речь, генеративная модель, реальное время.

# Вступ

Системи розпізнавання мовлення поступово займають місце посередника між людиною і комп'ютером, витісняючи звичні засоби введення інформації. Для англійської мови поруч із програмним забезпеченням диктування на ПК з'явився ряд мережних сервісів, що обслуговують введення голосом пошукових запитів або дають змогу диктувати лист електронної пошти [1]. При цьому демонструється доволі прийнятна працездатність таких систем, навіть враховуючи помітну затримку при користуванні мережними *cloud*-сервісами. Очевидно, що такі системи (а) – оперують доволі широким лексиконом і (б) – виконують обчислення в реальному часі.

Аналіз патентів комерційних фірм і публікацій провідних наукових центрів показує, що найбільш поширена у світі схема розпізнавання мовленнєвого сигналу в рамках генеративної моделі або прихованої (неявної) марківської моделі (Hidden Markov Model – HMM) побудована на генеруванні послідовності композитних мовленнєвих образів (слів або фраз), складених із акустичних моделей фонем, вже на етапі акустичного декодування [2], [3]. Одночасно, за лінгвістичною моделлю, оцінюється та враховується вірогідність гіпотетично розпізнаних послідовностей слів шляхом прогнозування поточного слова-претендента за одним або більше словами-попередниками.

Загальновідомо, що слов'янські мови характеризуються такими властивостями, як величезна кількість словоформ (у 8 – 10 разів більше, ніж в англійській мові) та відносно вільний порядок слів у реченні. Це призводить до стрімкого зростання робочого словника та до зменшення сили прогнозування в лінгвістичній моделі. Тому придатність загальноприйнятих методів і алгоритмів при розпізнаванні слов'янських мов підлягає сумніву, і це одна з причин пошуку нових схем розпізнавання, зокрема таких, що передбачають композицію слів за результатами акустичного декодування [4].

Сьогодні за допомогою систем розпізнавання мовлення ізольовано вимовлені слова та злите підготоване мовлення (наприклад, читання новин) розпізнається з надійністю близько 95% [1], [3]. У той же час розпізнавання спонтанного мовлення має набагато гірші результати. Розпізнавання спонтанного мовлення у реальних умовах спілкування (наприклад, за наявності шумів) є надзвичайно актуальною задачею, вирішення якої значно розширить сферу використання систем розпізнавання мовлення.

Вважаючи за необхідне продовжувати дослідження нових підходів, разом із тим стверджуємо, що достеменно не відомий резерв опрацьованої багатьма роками схеми розпізнавання [2], [3]. Адже досі не з'ясовано, наскільки системи на основі загальноприйнятого підходу обмежені в лексиці зі збереженням працездатності розпізнавання в реальному часі на обчислювальній базі, доступній пересічному користувачеві. Тому **ціль даної статті** – побудувати систему реального часу, яка може експлуатуватися на сучасному ПК для перетворення мовленнєвого сигналу на текст та як диктувальна машина.

У наступному розділі описуються засоби побудови бази даних і знань для розпізнавання мовлення, обґрунтовується вибір усних та писемних даних, приділяється увага перетворенню графем на фонеми, врахуванню ознак спонтанності. Далі описується діюча система, її характеристики та можливості застосування. У висновках пропонується ряд удосконалень, обговорюється сучасний стан досліджень та їх подальші перспективи.

# Параметри генеративної моделі та їх оцінювання

Вхідний мовленнєвий сигнал перетворюється на послідовність акустичних векторів фіксованого виміру $\mathbf{Y}_{1:T} = \left(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\right)$ у результаті препроцесингу. Тобто відбувається перехід у простір первинних ознак. Потім декодер намагається знайти послідовність слів $\mathbf{w}_{1:L} = \left(w_1, w_2, ..., w_L\right)$, яка найбільш вірогідно відповідає спостережуваному **Y**. Іншими словами, декодер має відшукати

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P\left(\mathbf{w} \mid \mathbf{Y}\right). \quad (1)$$

Не зважаючи на складність, ряд дискримінантних моделей намагається оперувати з цим виразом напряму [5]. Утім, найбільш продуктивною є генеративна модель, що розглядає еквівалентну задачу, яка виникає внаслідок застосування правила Баєса до (1):

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p\left(\mathbf{Y} \mid \mathbf{w}\right) P\left(\mathbf{w}\right). \quad (2)$$

Міра схожості $p\left(\mathbf{Y} \mid \mathbf{w}\right)$ становить акустичну складову, а ймовірність $P\left(\mathbf{w}\right)$ – лінгвістичну складову генеративної моделі розпізнавання мовленнєвого сигналу.

Розглянемо детальніше акустичну складову або **акустичну модель** (АМ). Кожне вимовлене слово $w$ розкладається на послідовність $L_w$ базових звуків, тобто фонем. Ця послідовність є вимовою слова або його фонемною транскрипцією $\mathbf{q}^{(w)}_{1:K_w} = \left(q_1, q_2, ..., q_{K_w}\right)$.

Під час розробки мовленнєвих технологій повинні враховуватися індивідуальні, ситуативні особливості мовлення диктора, вимова слів у потоці мовлення, а це спричиняє введення багатозначності при переході до фонемного тексту.

Щоб урахувати множинність варіантів вимови слова, міра схожості $p\left(\mathbf{Y} \mid \mathbf{w}\right)$ обчислюється за багатьма фонемними транскрипціями:

$$p\left(\mathbf{Y} \mid \mathbf{w}\right) = \sum_{\mathbf{Q}} p\left(\mathbf{Y} \mid \mathbf{Q}\right) P\left(\mathbf{Q} \mid \mathbf{w}\right). \quad (3)$$

У цьому виразі сума береться за всіма допустимими послідовностями вимови для **w**, **Q** – деяка послідовність фонемних транскрипцій, для якої виконується:

$$P\left(\mathbf{Q} \mid \mathbf{w}\right) = \prod_{l=1}^{L} P\left(\mathbf{q}^{(w_l)} \mid w_l\right), \quad (4)$$

де $\mathbf{q}^{(w_l)}$ – допустима вимова слова $w_l$.

На практиці, при обчисленні виразу (3) береться максимум замість суми, а за рахунок зменшення варіантів альтернативної вимови слів досягається економія ресурсів при обчисленні (4).

Акустична сутність фонеми $q$ подається у вигляді генеративної моделі, як показано на рис. 1а, де $\{a_{ij}\}$ – статистичні параметри переходу між станами, $\{b_j(\,)\}$ – розподіли у просторі первинних ознак для робочих станів.

Ці розподіли фактично апроксимують у просторі первинних ознак ті області, через які проходять траєкторії, що відповідають акустичній реалізації фонеми $q$. Такий загальний вигляд має базова НММ.

Технічно перехід від робочого стану генеративної моделі до одного зі станів, з яким робочий стан пов'язаний, здійснюється за одиницю відліку часу, а матриця $\{a_{ij}\}$ залежить від топології НММ та має вигляд стохастичної матриці, що формує ланцюг Маркова.

Допустима послідовність станів

$$\Theta_{1:T} = (\theta_1, \theta_2, \ldots, \theta_T), \tag{5}$$

за якою генерується еталонний (модельний) сигнал, є деякою акустичною транскрипцією спостережуваного сигналу.

Відповідно до генеративної моделі, ці стани пов'язані умовними залежностями як між собою, так і з відліками спостережуваного сигналу.

На рис. 1б ці залежності для базової НММ подані у вигляді динамічної баєсівської мережі (ДБМ) [3].

У прийнятій тут нотації дискретні змінні зображено в квадратах, неперервні змінні – у колах, спостережувані змінні затінені, а приховані – на світлому тлі.

Цей вигляд зручний для ілюстрації розширень базової генеративної моделі, зокрема для введення додаткових параметрів і залежностей, наприклад, між сусідніми відліками спостережуваного сигналу.

Крім того, ДБМ зручна для пояснення дискримінантних моделей.

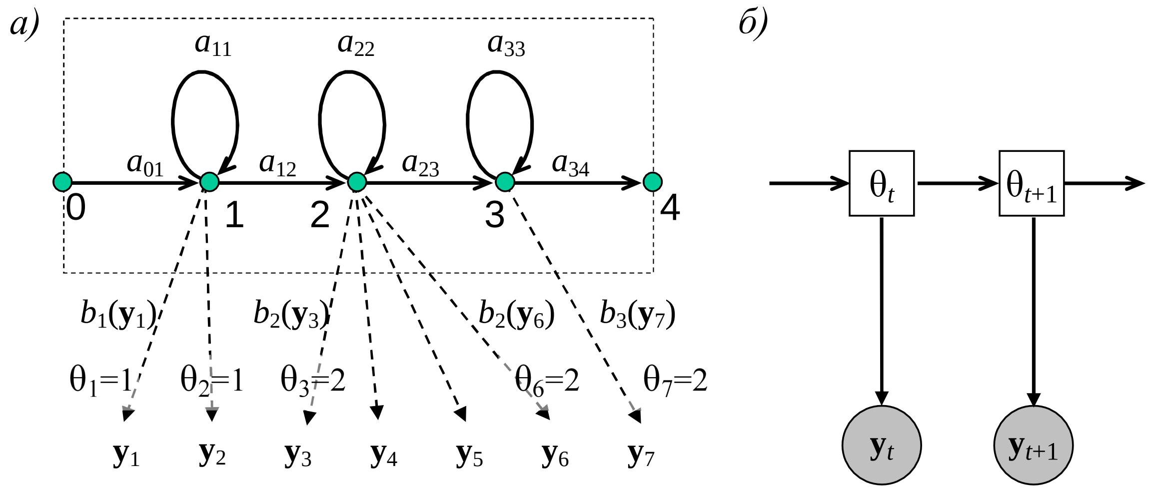



Рисунок 1 – Базова генеративна модель (НММ) фонеми: а) – у вигляді згорнутого графа динамічного програмування та б) – в термінах динамічної баєсівської мережі

Для кращої якості апроксимації областей перебування фонеми замість одного нормального закону (гаусоїда) $G(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$ вводиться суміш гаусоїдів:

$$b_j(\mathbf{y}) = \sum_{m=1}^{M} c_{jm} G\left(\mathbf{y};\boldsymbol{\mu}^{(jm)},\boldsymbol{\Sigma}^{(jm)}\right), \tag{6}$$

де $c_{jm}$ – апріорна ймовірність перебування у $m$-у гаусоїді $j$-о стану, яка задовольняє умовам функції ймовірності, зокрема $c_{jm} \geq 0$ і $\sum_{m=1}^{M} c_{jm} = 1$.

Сумішшю гаусоїдів моделюються асиметричні розподіли та розподіли з багатьма модами. Це дає змогу точніше відобразити розмаїття сигналу на акустичному рівні.

Важливим питанням є обґрунтоване забезпечення діагональності кожної коваріаційної матриці $\mathbf{\Sigma}^{(jm)}$. Для цього, при потребі, проводиться декореляція простору первинних ознак шляхом застосування дискретного косинус-перетворення. Таким чином, апроксимація областей перебування фонем здійснюватиметься об'єднанням еліпсоїдів, витягнутих уздовж осей координат.

На рис. 2 зображено проекцію на двовимірний простір траєкторії руху реалізації слова *оса́* у просторі первинних ознак. Відліки спостережуваного сигналу $\mathbf{y}_{t=1:72}$ проходять через області перебування відповідних фонем: # (фонема-пауза), *о*, *с*, *А* (а наголошена), #. Фонема-пауза # апроксимується еліпсоїдом, що відповідає одному гаусоїду в єдиному стані моделі цієї фонеми #1. Припускається, що ймовірність апроксимації гаусоїдом деякої точки всередині відповідного еліпсоїда більша за 0,1. Моделі фонем *о* та *А* містять по три стани: *о*1, *о*2, *о*3 та *А*1, *А*2, *А*3, розподіл кожного з них апроксимується двома компонентами суміші нормальних законів. Гаусоїди, що відповідають одному й тому ж стану в межах фонеми, мають однакове штрихування. Модель фонеми *с* містить також три стани, але для апроксимації розподілу кожного зі станів використовується лише один гаусоїд.

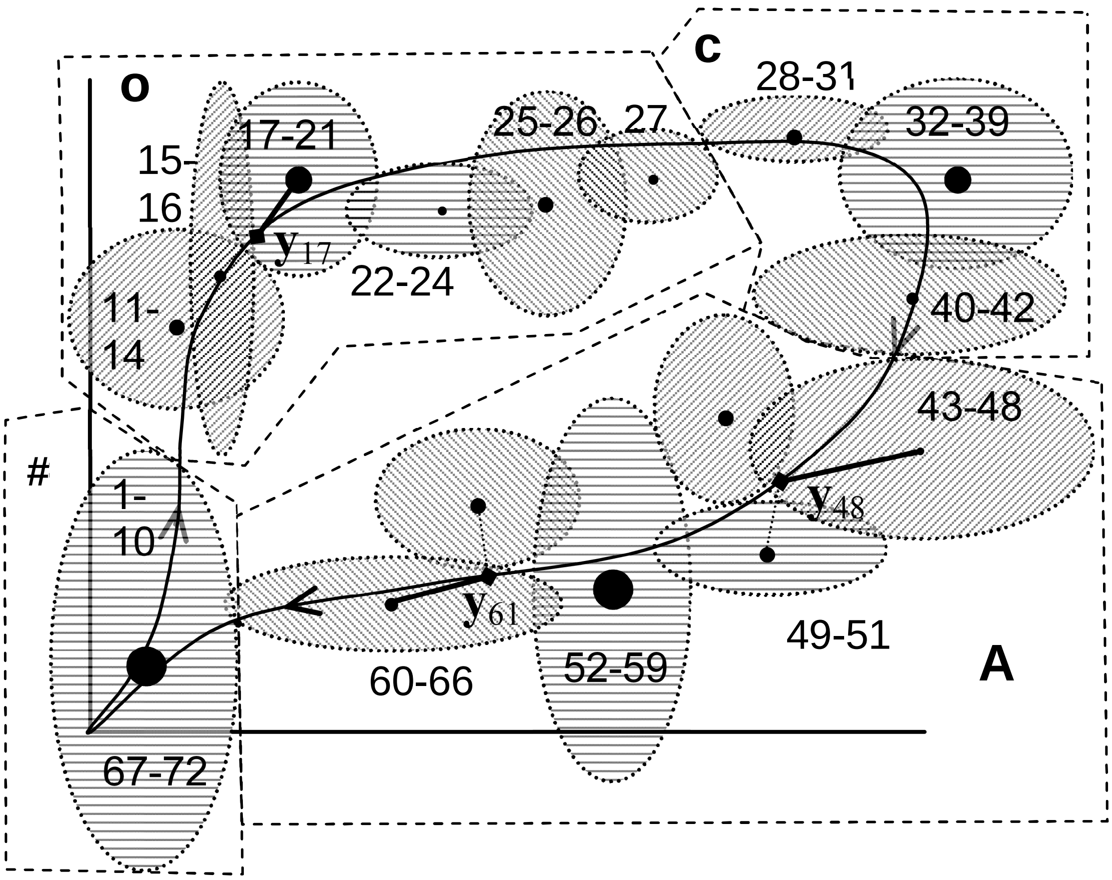



Рисунок 2 – Проекція на двовимірний простір траєкторії руху реалізації слова *оса́* у просторі первинних ознак

У процесі розпізнавання методом динамічного програмування серед усіх допустимих акустичних транскрипцій шукається така, що найкращим чином апроксимує траєкторію сигналу. Зображена на рис. 2 траєкторія сигналу найкраще апроксимується акустичною транскрипцією вигляду (5), що набуває значень: $\theta_{1:10} = \#1$, $\theta_{11:16} = o1$, $\theta_{17:24} = o2$, $\theta_{25:27} = o3$, $\theta_{28:31} = c1$, $\theta_{32:39} = c2$, $\theta_{40:42} = c3$, $\theta_{43:48} = A1$, $\theta_{49:59} = A2$, $\theta_{60:66} = A3$, та $\theta_{67:72} = \#1$.

Радіус чорного кола в точці математичного сподівання гаусоїда відповідає сукупно значенням ймовірності переходу в той же стан та апріорній імовірності перебування в гаусоїді цього стану згідно з (6). Числовий проміжок вказує ті часові відліки, які найкраще апроксимуються гаусоїдом. Для деяких гаусоїдів такий проміжок відсутній. Маркером квадратної форми на траєкторії показано окремі відліки. Центр гаусоїда, який найкраще цей відлік апроксимує, з’єднаний із ним суцільною лінією.

Параметри акустичної моделі оцінюються за мовленнєвим корпусом ітераційно. Спочатку вводиться одна компонента суміші нормального закону. Потім поступово нарощуються кількість гаусоїдів шляхом розщеплення тих, що мають найбільшу норму коваріаційної матриці. Максимальна кількість гаусоїдів оцінюється з розрахунку не менше 50 реалізацій фонеми на один гаусоїд.

**Лінгвістична складова** моделі (2) полягає в оцінюванні ймовірності

$$P(\mathbf{w}) = \prod_{k=1}^{K} P(w_k \mid w_{k-1}, \ldots, w_1). \tag{7}$$

Кількість попередніх слів може бути якої завгодно довжини, тому, з міркувань уможливлення реалізації обчислень, доцільно її обмежити до $N - 1$, і таким чином сформувати лінгвістичну модель (ЛМ):

$$P(\mathbf{w}) = \prod_{k=1}^{K} P(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_{k-N+1}), \tag{8}$$

де $N$ обирається в межах від 2 до 4. Імовірності $N$-грам оцінюються за текстовим корпусом шляхом статистичного підрахунку. Наприклад, якщо позначити через $C(w_{k-N+1}, \ldots, w_{k-1}, w_k)$ частоту $N$-грами $(w_{k-N+1}, \ldots, w_{k-1}, w_k)$, то

$$P(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_{k-N+1}) \approx \frac{C(w_{k-N+1}, \ldots, w_{k-1}, w_k)}{C(w_{k-N+1}, \ldots, w_{k-1})}. \tag{9}$$

Найбільшою теоретичною проблемою при побудові ЛМ є оцінка ймовірностей тих $N$-грам, для яких не набирається достатньо статистики. Тоді ця оцінка проводиться на підставі статистик $(N - 1)$-грам [3]. Іншою проблемою є наявність у текстовому корпусі слів, які не ввійшли до робочого словника. Прийнятним вирішенням цієї проблеми є введення категорії *невідомого слова*, що замінює в текстовому корпусі всі позасловникові слова. Крім того, значні фізичні обсяги ЛМ можуть стати на перешкоді практичного використання системи розпізнавання.

# Побудова діючої системи та її дослідна експлуатація

На рис. 3 зображено загальну структуру базової системи перетворення мовленнєвого сигналу на текст, що має компоненту реального часу, у якій реалізовано власне декодер, та компоненту, яка у відкладеному режимі здійснює оцінювання

параметрів математичної моделі. Для створення базової системи використано як власні розробки, так і різноманітний програмний інструментарій доступний в Інтернеті: *HTK, HTS, Julius, MITLM, CMU LM* [5, 6, 7].

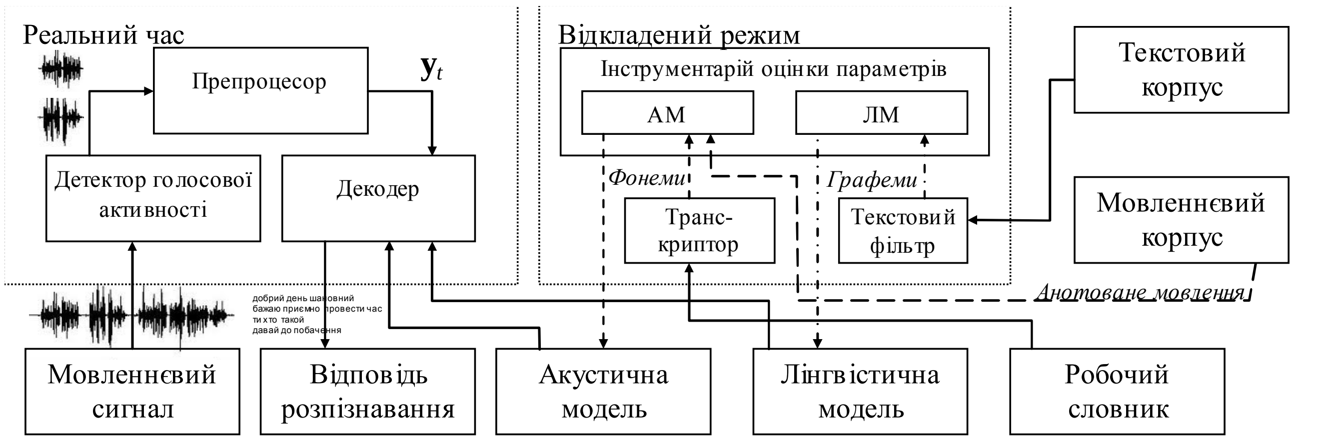



Рисунок 3 – Загальна структура базової системи перетворення мовленнєвого сигналу на текст

*Компонента реального часу* отримує *мовленнєвий сигнал* через одне з доступних джерел (мікрофон або файл). При проходженні через *детектор голосової активності* сигнал розбивається на сегменти за ознаками наявності голосового введення. Використовуються прості ознаки в амплітудно-часовому просторі на основі амплітуди та кількості переходів через нуль. *Блок препроцесора* переводить сигнал у простір первинних ознак. При цьому застосовано мел-кепстральне перетворення з відніманням середнього значення. *Декодер* порівнює вхідний сегмент із гіпотезами еталонного сигналу відповідно до (2) – (8), застосовуючи деяку обережну стратегію відкидання мало перспективних гіпотез [6]. Для цього використовується акустична та лінгвістична складові математичної моделі. Послідовність слів, яка генерує найбільш схожий еталонний сигнал, оголошується *відповіддю розпізнавання*.

*Акустичну модель* сформовано на основі однієї з перших версій мовленнєвого корпусу АКУЕМ [8]. Ця версія корпусу містила менше 40 годин розмічених експертами звукових записів українського мовлення (помилки анотації складали близько 5 – 6%). Топологія HMM кожної фонеми відповідає рис. 1а, за винятком фонеми-паузи, що допускає перехід із 3-о стану в 1-й, та короткої паузи, яка містить лише один робочий стан та допускає його пропуск. На відміну від рекомендацій [5], уточнювання параметрів робочого стану короткої паузи проводиться незалежно від фонеми-паузи. Нарощування гаусоїдів відбувається поступово, з більшою швидкістю для частотних фонем. Максимальна кількість гаусоїдів у стані фонеми – 36.

*Робочий словник* системи розпізнавання складається із частотного словника текстового корпусу та додаткових словників (словники соціальних і територіальних діалектів, словник суржику, словники власних назв, абревіатур та ін.). На відміну від англійської, для української мови до алфавіту фонем включено як наголошені, так і ненаголошені голосні. Інформація про місце наголосу у словах отримується із словника УМІФ [9], для додаткових словників наголоси проставляються експертом або прогнозуються []. Найчастотніші одно- та двоскладові слова доповнені варіантом без наголосу.

*Транскриптор* перетворює слова з інформацією про наголос на послідовність фонем, за якими створюються композитні акустичні моделі слів як для декодера, так і при оцінюванні параметрів АМ. У транскриптор закладено багатозначний підхід [],

який моделює особливості, властиві конкретній мові, на основі вивчення закономірностей зв'язку між орфографічними символами та символами з алфавіту фонем. Експертом сформовані правила перетворення «графема–фонема», у яких передаються індивідуальні особливості вимови дикторів і частково враховано коартикуляцію та редукцію звуків у потоці мовлення. У середньому на кожне слово припадає не більше 1, 2 варіанта транскрипції. Також вирішується проблема розшифрування чисел і символів.

В основу *текстового корпусу* для лінгвістичної моделі покладено матеріал, завантажений з ряду Інтернет-сайтів, що містять тексти новин та публіцистики (60%), художніх творів (8%), енциклопедичного характеру (24%), текстів юридичного спрямування (8%).

Потрібно зазначити, що серед матеріалу, завантаженого з сайтів новин, містяться коментарі та відгуки відвідувачів, тобто присутні текстові зразки спонтанного типу мовлення.

Під час оброблення текстового корпусу *текстовим фільтром* числа та символи перетворювалися на слова.

Було вилучено зайві фрагменти, повтори на рівні абзаців, речення, що містять суттєвий відсоток слів, відсутніх у словнику УМІФ. Загальний обсяг текстового корпусу складає 2 ГБ, куди ввійшло 17,5 млн речень або біля 250 млн реалізацій слів.

Оброблений текст надходить на вхід інструментарію формування *лінгвістичної моделі* на основі *N*-грам.

При цьому додатково вилучаються речення, які містять певний відсоток слів, відсутніх у робочому словнику, а у реченнях, що залишаються, такі слова позначаються як *невідомі*.

Максимальний порядок сформованої моделі – 3. Для робочого словника на 100 тисяч слів загальна кількість 3-грам становить 88,5 мільйонів, частка невідомих слів склала близько 2,5%, обсяг файлу – 1,2 ГБ.

Для моделювання ознак спонтанного мовлення введено клас *прозорих* слів, куди ввійшли екстралінгвістичні явища (неінформативні слова та звуки).

На основі компоненти реального часу (рис. 3) розроблено базову систему перетворення мовленнєвого сигналу на текст, що використовується для експериментальних досліджень. Графічний інтерфейс користувача, доданий до базової системи (рис. 4), дає змогу демонструвати розпізнавання злитого мовлення в реальному часі на ПК [**Ошибка! Источник ссылки не найден.**0].

Умови експлуатації розробленої системи враховують очікування потенційного користувача.

Словник системи покриває загальновживану лексику та множину слів деяких предметних областей: наприклад, природничі науки, будівництво, медицина, юриспруденція тощо.

У нашому випадку обрано тематику новин (політика, економіка, культура, спорт і погода). На акустичному рівні, система сприймає мовлення будь-якого адекватного користувача.

Заздалегідь підготоване мовлення, прочитані тексти, спонтанні висловлювання розпізнаються на одному рівні.

Щодо вимог до якості запису мовленнєвого сигналу доступними для пересічного громадянина засобами, не розглядаються сильно зашумлені записи та перекриття мовлення різних осіб в одному каналі запису.

Під час дослідної експлуатації цієї системи використовувалися словники на 10, 20, 50 і 100 тисяч слів. Оскільки для всіх словників декодування відбувалося в

реальному часі (до 15% на процесорі *i*7), було проведено більш детальне дослідження максимального словника у 100 тисяч слів.

Система тестувалася як диктувальна машина десятьма експертами. В умовах експлуатації, описаних вище, послівна помилка розпізнавання становить у середньому 10%. Перевірено ефективність поповнення словника новими словами, що сприймаються як *незнайомі* на рівні лінгвістичної моделі.

Експертами у словник додавалися власні назви та рідкісна термінологія. Можливість ставити голосом розділові знаки, починати новий абзац та відміняти останню операцію (у формі голосової команди, виділеної паузами) підвищила ергономіку системи в цілому.

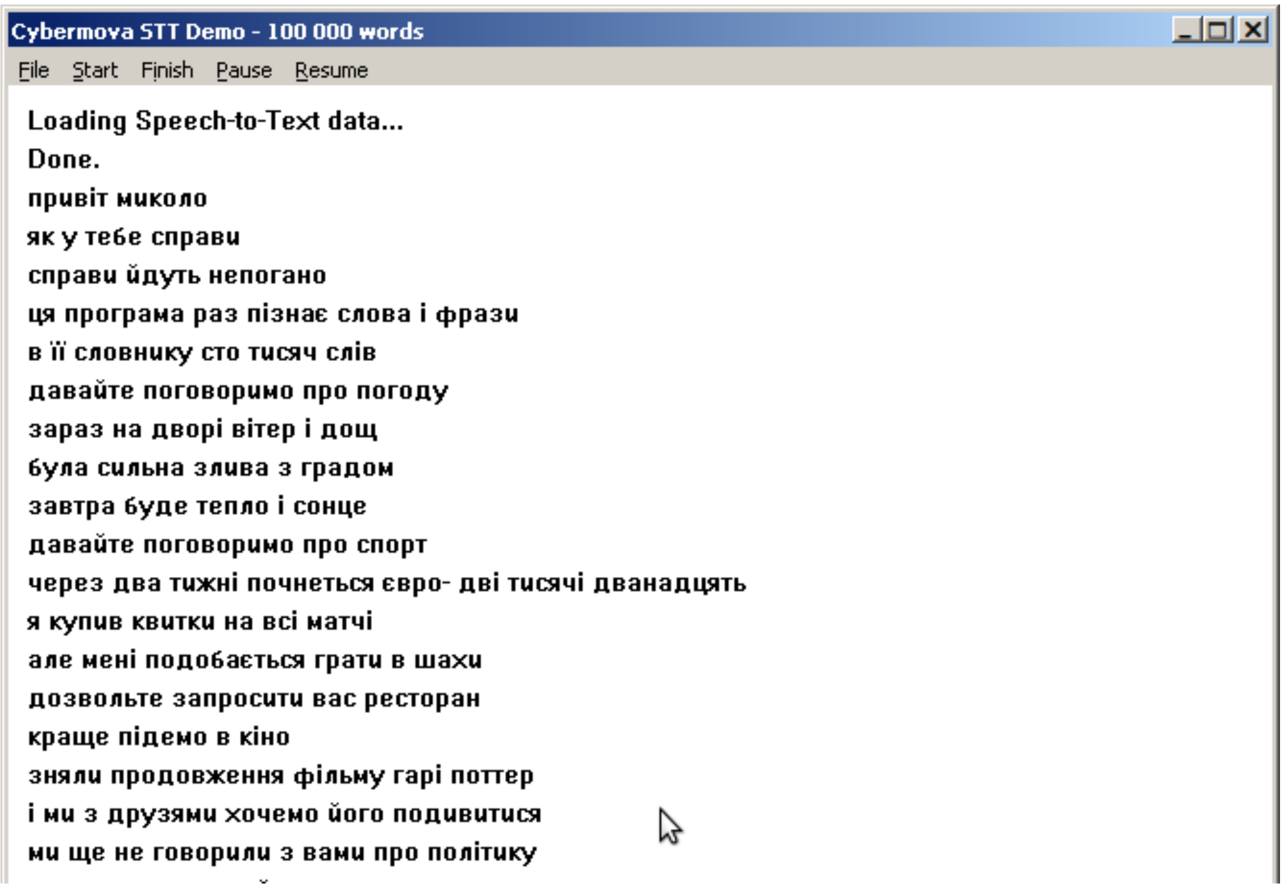



Рисунок 4 – Система диктування на ПК демонструє послівну помилку розпізнавання менше 5% на фрагменті з 90 слів загальної лексики

# Висновки

У статті досліджено найбільш поширену у світі схему розпізнавання мовленнєвого сигналу, що реалізує принцип *аналіз через синтез*. Створена на основі цієї схеми система перетворення мовлення на текст демонструє прийнятну працездатність при дослідній експлуатації.

Робота над описаною системою розпізнавання перебуває в початковій стадії. У найближчому майбутньому передбачається здійснити ряд заходів, що покращать надійність розпізнавання та розширять сферу використання системи. Ці заходи стосуються збільшення словника, оптимізації лінгвістичної моделі шляхом уведення класів слів, застосування контекстно залежних моделей фонем, кластеризацію дикторів та настроювання на голос диктора, передбачення знаків пунктуації та регістру слів.

Важливим завданням залишається суттєве розширення бази навчальної вибірки для акустичної та лінгвістичної компонент моделі. Цьому сприятиме вирішення задачі відповідності тексту і сегмента мовленнєвого сигналу. Актуальним залишається більш точне перетворення чисел і символів на графеми, зокрема з урахуванням роду й відмінків та їх неоднозначності.

Для систем диктування не менш важливо розвинути взаємодію з користувачем при редагуванні тексту: пропонувати варіанти виправлення, використовуючи багатозначність відповіді розпізнавання, та запам'ятовувати виправлення при подальшому диктуванні. Потрібно передбачити розширення робочого словника користувачем через віднесення нових слів до категорії *невідомого слова*, а також через оновлення параметрів лінгвістичної моделі.

Для поліпшення результатів розпізнавання планується посилити відповідність лінгвістичної моделі предметній області, стилю та жанру мовлення. Для досягнення цього, текстовий корпус лінгвістичної моделі потрібно розбити на декілька частин та провести їх інтерполяцію з метою мінімізувати ентропію для зразків текстів потрібної предметної області.

# Література

1. [Електронний ресурс]. – Режим доступу : http://www.forbes.com/sites/ greatspeculations/2011/11/15/apple-trumps-google-on-voice-recognition-in-head-to-head-test/
2. Винцюк Т.К. Анализ, распознавание и смысловая интерпретация речевых сигналов / Винцюк Т.К. – Киев : Наукова думка, 1987. – 264 с.
3. Gales M. The Application of Hidden Markov Models in Speech Recognition / M. Gales, S. Young // Foundations and Trends in Signal Processing, 2007. – № 1(3). – P. 195-304.
4. Taras Vintsiuk. Multi-Level Multi-Decision Models for ASR / Taras Vintsiuk, Mykola Sazhok // Proceedings of the 10th Int. Conference on Speech and Computer. – SpeCom'2005, Patras, 2005. – P. 69-76.
5. Young S.J. The HTK Book Version 3.4 / S.J. Young [et al.]. – Cambridge University, 2006.
6. Lee A. Recent Development of Open-Source Speech Recognition Engine Julius / A. Lee, T. Kawahara // Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.
7. Bo-June (Paul) Hsu. Iterative Language Model Estimation : Efficient Data Structure & Algorithms / Bo-June (Paul) Hsu, James Glass // In Proc. Interspeech, 2008.
8. Васильєва Н.Б. Створення акустичного корпусу українського ефірного мовлення / Н.Б. Васильєва, В.В. Пилипенко, О.М. Радуцький,// Обробка сигналів і зображень та розпізнавання образів : X Міжнар. конференція : «УкрОбраз'2010». – Київ, 2010. – С. 55-58.
9. Широков В.А. Організація ресурсів національної словникової бази / В.А. Широков, В.В. Манако // Мовознавство. – 2001. – № 5. – С. 3-3.
10. [Електронний ресурс]. – Режим доступу : www.cybermova.com/products/stt-demo.htm
11. Gales M. Discriminative models for speech recognition / M. Gales // ITA Work-shop. – University San Diego, USA. – February, 2007.Електронний ресурс].
12. Zweig G. Speech Recognition with Dynamic Bayesian Networks / G. Zweig // PhD-thesis. – University of California, Berkeley. – 1998.
13. Робейко В.В. Багатозначна багаторівнева модель перетворення орфографічного тексту на фонемний / В.В. Робейко, М.М. Сажок // Штучний інтелект. – Донецьк, 2011. – № 4. – С. 117-125.
14. Робейко В.В. Використання текстового корпусу для прогнозування наголосів у словах української мови / В.В. Робейко, М.М. Сажок // Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту : матеріали міжнародної наукової конференції. – Херсон, 2012. – С. 171-172.

# Literatura

1. http://www.forbes.com/sites/greatspeculations/2011/11/15/apple-trumps-google-on-voice-recognition-in-head-to-head-test/
2. Vintsiuk T.K. Analiz, raspoznavaniye i smyslovaya interpretatsiya rechevykh signalov. – Kiev : Naukova Dumka, 1987. – 264 p.

3. M. Gales and S. Young. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing. – 2007. – № 1(3). – P. 195- 304.
4. Taras Vintsiuk, Mykola Sazhok. Multi-Level Multi-Decision Models for ASR // Proceedings of the 10th Int. Conference on Speech and Computer. – SpeCom’2005, Patras, 2005. – P. 69-76.
5. Young S.J. et al., The HTK Book Version 3.4, Cambridge University, 2006.
6. A. Lee, T. Kawahara. "Recent Development of Open-Source Speech Recognition Engine Julius" APSIPA ASC, 2009.
7. Bo-June (Paul) Hsu and James Glass. Iterative Language Model Estimation: Efficient Data Structure & Algorithms. In Proc. Interspeech, 2008.
8. N.B. Vasyliyeva, V.V. Pylypenko, O.M. Raduts'kyy, V. Robeiko, M. Sazhok. Stvorennia akustychnoho korpusu ukrayins'koho efirnoho movlennia // UkrObraz’2010 – P. 55-58.
9. Shyrokov V.A., Manako V.V. Orhanizatsiya resursiv natsional'noyi slovnykovoyi bazy // Movoznavstvo. – № 5. – 2001 – P. 3-13.
10. www.cybermova.com/products/stt-demo.htm G. Zweig. Speech Recognition with Dynamic Bayesian Networks. PhD thesis, UC Berkeley, 1998.
11. M. Gales. Discriminative models for speech recognition // ITA Work-shop, University San Diego, USA, February 2007.
12. Zweig G. Speech Recognition with Dynamic Bayesian Networks / G. Zweig // PhD-thesis. – University of California, Berkeley. – 1998.
13. V. Robeiko, M. Sazhok. Bahatoznachna bahatorivneva model' peretvorennia orfohrafichnoho tekstu na fonemnyy. Shtuchnyy intelekt. – Donets’k, 2011. – № 4. – P. 117-125.
14. V. Robeiko, M. Sazhok. Vykorystannia tekstovoho korpusu dlia prohnozuvannia naholosiv u slovakh ukrayins'koyi movy // ISDMCI’2012. – P. 171-172.

## *RESUME*

***V.V. Robeiko, M.M. Sazhok***

## Real-Time Spontaneous Speech Recognition Based on Word Acoustic Composite Models

This paper describes implementation of methods and algorithms for the automatic speech recognition based on word composition proceeding from acoustic phoneme models. Such a design of the speech-to-text decoder is conventional throughout the world and is most productive for Western languages [3]. The aim is to explore the conventional speech recognition approach applied to the Ukrainian language.

Comparatively to Western languages, Slavonic languages like Ukrainian are highly inflective with relatively free word order. This means that the working dictionary grows in times and perplexity of the language model is huge enormously. But till now no one answered how restricted must be a conventional speech recognition system to have an acceptable performance in real time on a modern PC. To answer this question experimentally we use own and widely available toolkits for speech and language processing.

Firstly, we analyze the data-driven methods to estimate parameters for both acoustic and linguistic components of the mathematical model. 40 hours of speech data are taken from the AKUEM corpus [84] to estimate HMM parameters for Ukrainian phonemes. 2 GB of downloaded and processed text data are converted to 3-gram language model. The grapheme-to-phoneme conversion procedure takes into account word stress issue and spontaneous continuous speech features [].

The basic experimental speech-to-text system is able to operate a 100k vocabulary occupying less than 15% of i7 processor time. Restricting the input speech to common lexica and media domain we may conclude the practical applicability of the system. A demo-version of the dictation machine is available for its performance appraisal [9]. Finally, we discuss the prospective of dictionary and domain extension, parameter estimation improvement and ergonomic issues.

*Стаття надійшла до редакції 03.07.2012.*